

Assignment 3

(Unsupervised Learning and Dimensionality Reduction)

Approach:

We shall be using the same two datasets as used in Assignment 1 (Supervised Learning). But we shall be using these datasets to see how the various Un-Supervised Clustering techniques are able to categorize the instances into clusters without the knowledge of their corresponding Class values (which is the ground truth labels). We shall also do a comparative analysis to see how the Dimensionality Reduction techniques affect the output of the Clustering techniques.

We shall use various scoring methods like Silhouette Score, BIC (Bayesian Information Criterion), Adjusted Mutual Information (AMI) to measure the quality of clusters as applicable.

Description of the two Datasets:**1. Dataset A: Breast Cancer Wisconsin (Diagnostic) Dataset**

This breast cancer dataset is obtained from UCI Machine Learning Repository and used as the first dataset (Dataset A). This dataset falls under the Binary Classification type problem where the target attribute (i.e., class) has two categorical values and contains 30 attributes (features).

The distribution of the 2 classes among the 569 instances are as follows. This shows that the distribution of the 2 class values is well balanced.

Benign: 357 (62.74%); Malignant: 212 (37.26%)

2. Dataset B: Statlog (Vehicle Silhouettes) Dataset

This vehicle silhouettes dataset is obtained from UCI Machine Learning Repository and used as the second dataset (Dataset B). The dataset contains 946 instances with 18 attributes (features).

The distribution of the 4 classes among the 946 instances are as follows. This shows that the distribution of the 4 class values is well balanced.

Opel: 240 (25.37%); Saab: 240 (25.37%); Bus: 240 (25.37%); Van: 226 (23.89%)

Reason of choosing these 2 datasets:

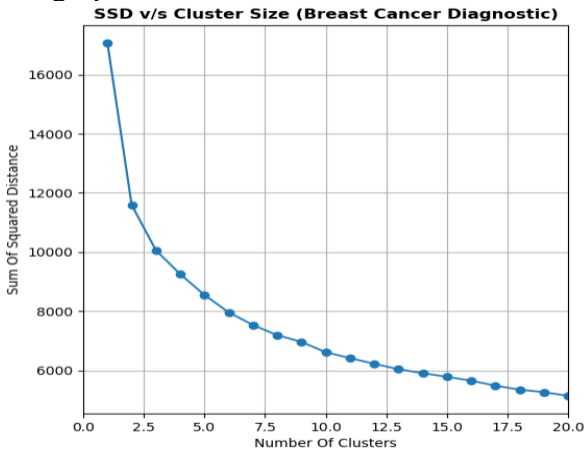
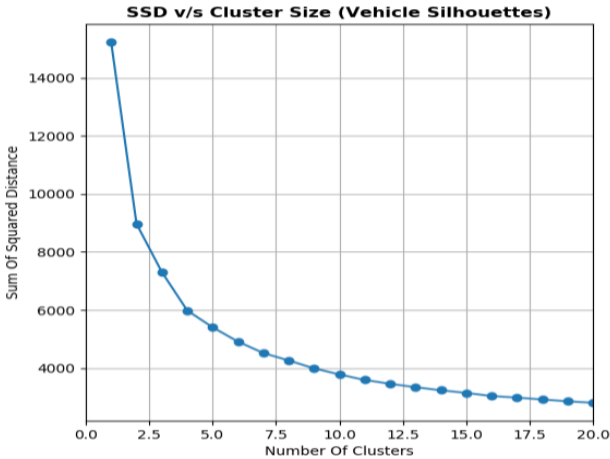
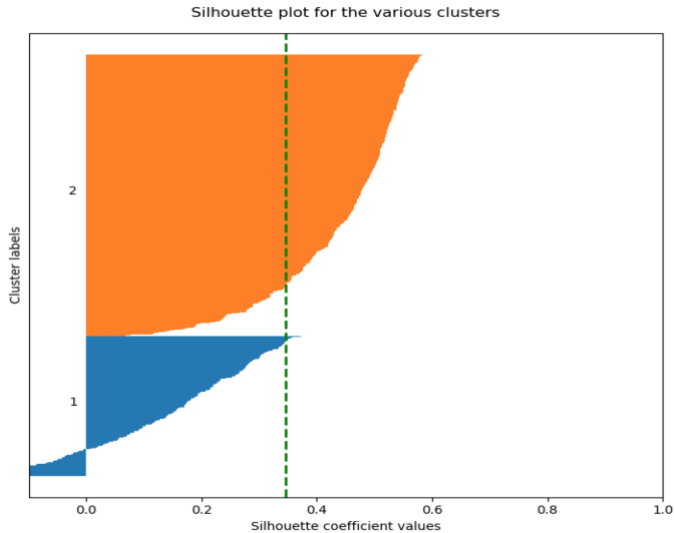
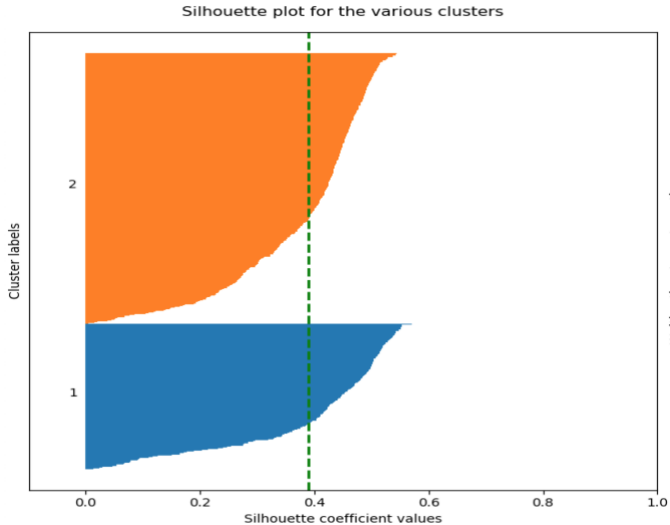
1. The above 2 datasets represent 2 different groups of Classification problems – Binary and Multi Classification.
2. There is an advantage of using datasets with known labels as we can use evaluation metrics like “Adjusted Mutual Information (AMI)” which is based on comparison with the Ground-Truth Labels.
3. The known number of class values in these datasets provides a baseline to see how different un-supervised clustering results are or how similar.
4. But we need to be careful of not using the known labels while fitting the models and should pretend to be completely unaware of them.

◇ Clustering (Without Dimensionality Reduction)

K-Means Clustering

Identify the optimal number of clusters with the Elbow method on the 2 datasets. In Elbow method we plot the sum of squared distance (SSD) between data points and their respective assigned centroids against each value of k . We shall select that value of k where SSD starts to flatten out forming an elbow. [1]

Silhouette Score shall be used to measure the quality of clusters. This metrics shows the degree of separation between clusters. So higher the value, better the quality of cluster. [1]

Dataset A (Breast Cancer)	Dataset B (Vehicle Silhouettes)
<p>We can see that at $k=2$ the elbow starts to form and the graph flattens out.</p> 	<p>Here also, at $k=2$ the elbow starts (momentum slows down) to form and the graph flattens out.</p> 
<p>$k=2$ also shows the max average Silhouette score Of 0.3434 where each cluster score above the average.</p> <p>$n_clusters = 2$, Avg. silhouette_score: 0.3434</p> 	<p>For $k=2$ shows the max average Silhouette score Of 0.3896 where each cluster score is above the average and looks pretty balanced.</p> <p>$n_clusters = 2$, Avg. Silhouette Score: 0.3896</p> 

Expectation Maximization with GMM Clustering

GMM treats the data as a overlapping placement of multiple Gaussian datasets with separate mean and variances. It then applies the Expectation-Maximization (EM) algorithm to approximation of mean and variances approximately. [2]

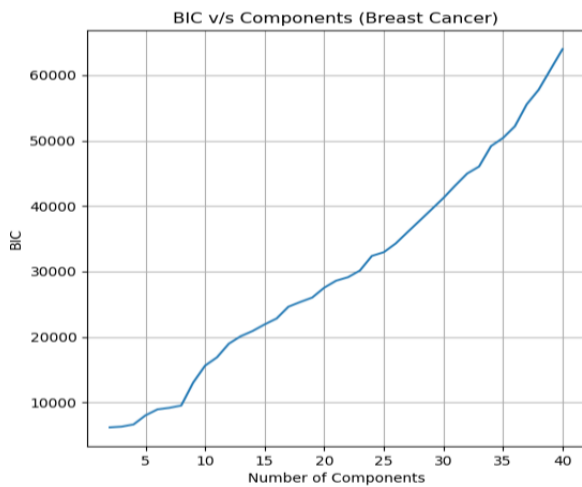
Silhouette scores works best with spherical clusters. Since GMM may not always produce spherical clusters, this metrics may not work best in this case.

That's why we shall use BIC (Bayesian Information Criterion) to select the optimal number of components. BIC penalizes if the number of Gaussians are large i.e., penalises any complex model (follows Occam's razor).[2]

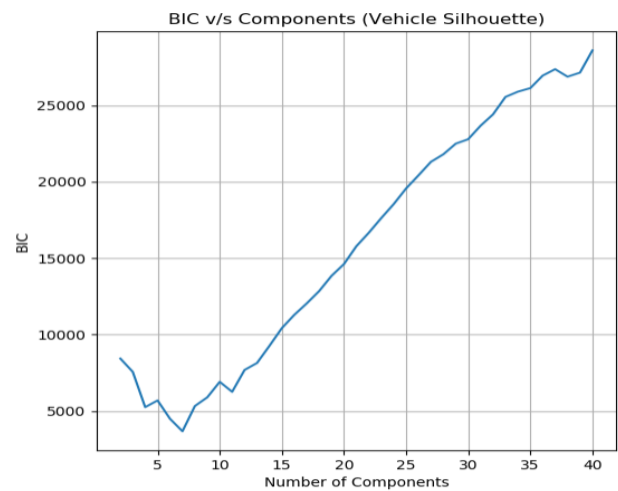
Since clusters with lower BIC values are preferred, for simplicity, we shall be selecting the cluster components with the lowest BIC score.

Dataset A (Breast Cancer)	Dataset B (Vehicle Silhouettes)
---------------------------	---------------------------------

Since we are choosing the optimal clusters based on lowest BIC score. Here optimal $k = 2$



Since we are choosing the optimal clusters based on lowest BIC score. Here optimal $k = 7$



◇ Dimensionality Reduction:

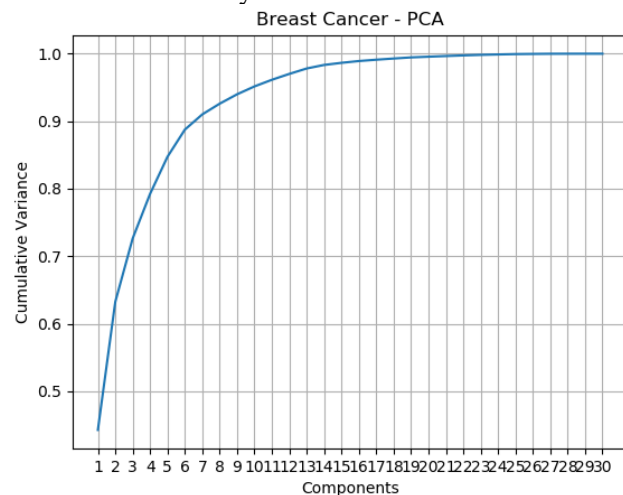
Here we shall be applying the following Dimensionality Reduction techniques on each of the datasets and re-apply the 2 clustering techniques. Then we shall do a comparative analysis of the clusters with before and after applying the Dimensionality Reduction techniques.

Principal Component Analysis (PCA):

Data Preprocessing: Data scaling shall be done using Python API on both the datasets to standardize the Feature values.

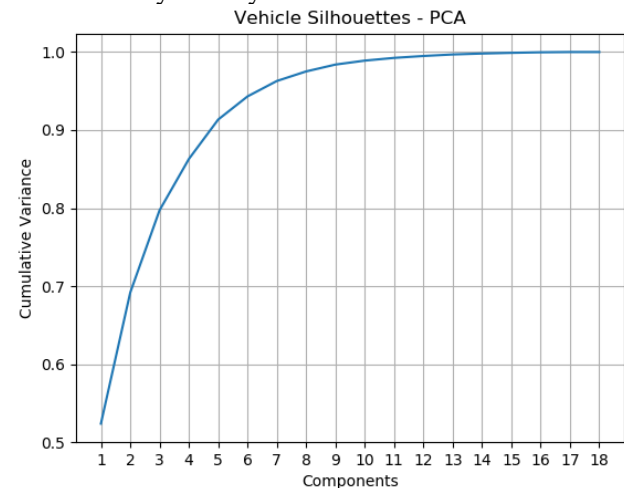
Dataset A: Breast Cancer

Based on the graph below, we can capture high variance of 0.9833 with only 14 feature components out of 30 i.e., Feature reduction by 53.3%.



Dataset B: Vehicle Silhouettes

Based on the graph below, a variance as high as 0.9838 with only 9 feature components out of 18 i.e., Feature Reduction by exactly 50%.



Re-running K-Means:

Using the Elbow Method, K-means shows the number of clusters to be the same and before running PCA i.e., $k=2$, with highest average Silhouette Score of 0.3491.

Re-running K-Means:

Using the Elbow Method, K-means shows the number of clusters to be the same and before running PCA i.e., $k=2$, with highest average Silhouette Score of 0.3491.

Re-Running EM-GMM:

Based on the lowest BIC score the optimal number of clusters is now k=4. This is different from k=2 before applying PCA.

Re-Running EM (GMM):

Based on the lowest BIC score the optimal number of clusters has now increased to k=9.

PCA always projects data into a lower dimensional space. This is the reason we can see that the AMI and Silhouette scores are different after applying dimensionality reduction technique.

Performance of Clustering Before/After PCA:

Optimal cluster selections are highlighted in Blue.

	Dataset A: Breast Cancer		Dataset B: Vehicle Silhouette	
	K-Means Silhouette Score	EM-GMM BIC Score	K-Means Silhouette Score	EM-GMM BIC Score
Before PCA	0.3434 (K=2) AMI Score = 0.5318	6174.7942 (K=2) AMI Score = 0.6538	0.3896 (K=2) AMI Score = 0.1225	3668.4657 (K=7) AMI Score = 0.2772
After PCA	0.3491 (K=2) AMI Score = 0.5540	20372.1102 (K=4) AMI Score = 0.4885 20552.6195 (K=2) AMI Score = 0.3087	0.3962 (K=2) AMI Score = 0.1225	16910.4532 (K=9) AMI Score = 0.2711 17171.7003 (K=7) AMI Score = 0.1558

For the k-Means clustering the Silhouette coeff. shows some increase after performing PCA. This shows that the cluster quality for k-Means have improved after doing PCA, probably due to removal of noisy attributes due to sensitivity of k-means to possible outliers.

For Expectation Maximization using GMM, the BIC scores for both the datasets have considerably gone up for both the datasets. This means that the quality of clusters for EM-GMM has gone worse after applying PCA.

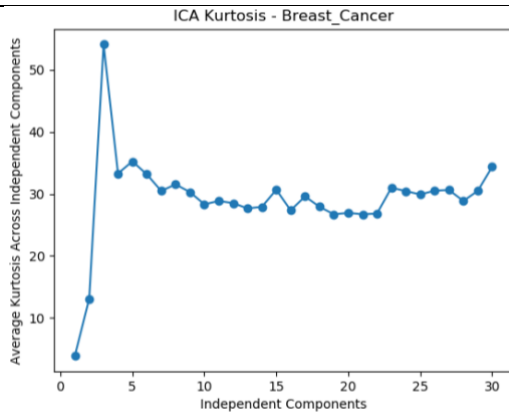
The AMI scores are used to validate how well the clusters line-up with the ground-truth labels.

1. Before applying PCA: AMI scores using Expectation Maximization (GMM) is higher than that of K-Means. Which means that clusters created by EM-GMM are better aligned with the ground-truth labels than k-means.
2. After applying PCA: PCA looks to have improved the performance of the K-Means clustering as AMI score have gone up. Whereas for Expectation Maximization (GMM) the AMI score fell below showing larger gap created with the real values (ground-truth labels). This may be due to reduced separability of PCA using hyperplanes.

Independent Component Analysis (ICA):**Data Preprocessing for ICA:**

1. Centering: Subtracting the mean [$X \leftarrow (X - X_{\text{mean}})$]
2. Whitening: With this process the potential correlations in between the features are removed (i.e., covariance becomes 0) and the variance of each of the components become 1.
3. The above-mentioned data preprocessing is already handled by Scikit-Learn API (FastICA(whiten=True)).

Dataset A: Breast Cancer	Dataset B: Vehicle Silhouettes
--------------------------	--------------------------------



We shall select the independent components (= 3) with highest Average Kurtosis value.

Re-running K-Means:

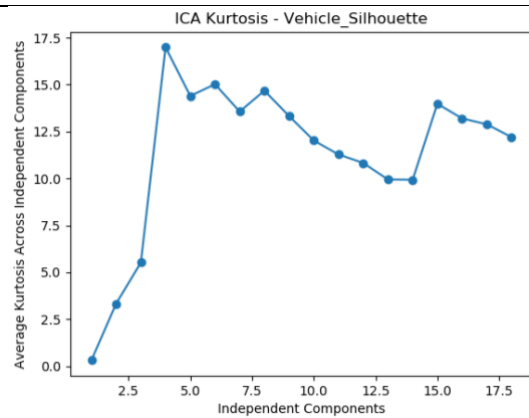
Using the Elbow Method, K-means shows the number of optimal clusters to be different (i.e., $k=4$).

But as per highest Silhouette_Score and AMI score, $k=2$ is still the optimal cluster size.

So, we shall select $k=2$ as optimal based on Silhouette and AMI score

Re-Running EM-GMM:

Based on the lowest BIC score (-8000.81) the optimal number of clusters is now $k=4$. This is different from $k=2$ before applying ICA.



We shall select the independent components (= 4) with highest Average Kurtosis value.

Re-running K-Means:

Using the Elbow Method, K-means shows the number of optimal clusters to be different (i.e., $k=6$).

The elbow method conforms with the highest Silhouette Score of 0.3391.

Re-Running EM (GMM):

Based on the lowest BIC score of -15438.15 the optimal number of clusters has now increased to $k=10$.

Performance of Clustering Before/After ICA:

Optimal cluster selections are highlighted in Blue.

	Dataset A: Breast Cancer		Dataset B: Vehicle Silhouette	
	K-Means Silhouette Score	EM-GMM BIC Score	K-Means Silhouette Score	EM-GMM BIC Score
Before ICA	0.3434 (K=2) AMI Score = 0.5318	6174.7942 (K=2) AMI Score = 0.6538	0.3896 (K=2) AMI Score = 0.1225	3668.4657 (K=7) AMI Score = 0.2772
After ICA	0.5034 (K=2) AMI Score = 0.4518	-8000.8084 (K=4) AMI Score = 0.5252	0.3391 (K=6) AMI Score = 0.1392	-15438.1524 (K=10) AMI Score = 0.2344
		-7713.6054 (K=2) AMI Score = 0.5525	0.2576 (K=2) AMI Score = 0.1422	-15243.1976 (K=7) AMI Score = 0.1698

Dataset A:

For the k-Means clustering the Silhouette coeff. shows some increase after performing ICA. This shows that the cluster quality for k-Means have improved after Feature Reduction.

EM-GMM also shows a very low BIC score. This show that after applying ICA the quality of clusters has improved. Where-as the AMI scores have gone down showing that the overall degradation of alignment of the clusters with the ground-truth labels.

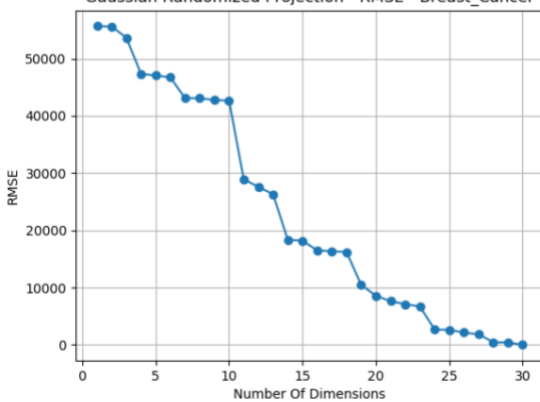
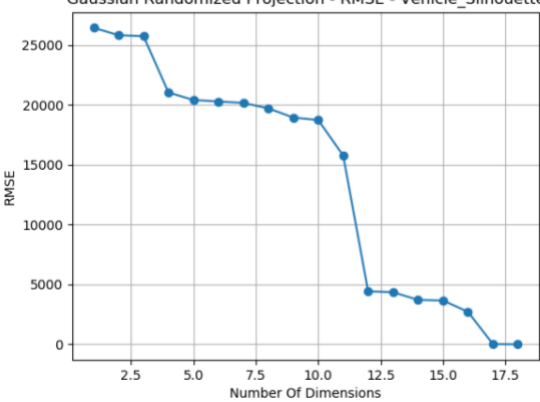
Dataset B:

For k-means the Silhouette coeff. have gone down slightly meaning a small degradation in cluster quality. The AMI has improved by a very small number showing a very small betterment in the cluster alignment with the ground-truth labels (at least no degradation).

EM-GMM shows a very low BIC score. This shows that after applying ICA the quality of clusters has improved by separating into 10 clusters with only 4 features. This shows the power of ICA. But the AMI score has gone slightly down meaning little degradation in alignment with ground-truth labels.

Randomized Projection (RP):

Since both the datasets are dense with no null or missing values, we shall be using Gaussian Random Projection technique.

Dataset A: Breast Cancer	Dataset B: Vehicle Silhouettes
<p>Gaussian Randomized Projection - RMSE - Breast_Cancer</p>  <p>The above RMSE plot shows a linear step pattern where the steepness of the curve slows down from 14. The reduction of RMSE is more (steeper slope) up to dimensions=14. Hence we shall select the number of dimensions as 14.</p> <p><u>Re-running K-Means:</u> Using the Elbow Method, K-means shows the number of optimal clusters to be different (i.e., k=2). This also conforms with the highest Silhouette_Score of 0.6991. So, we shall select k=2 as optimal cluster size.</p> <p><u>Re-Running EM-GMM:</u> Based on the lowest BIC score (16140.6673) the optimal number of clusters is now k=5.</p>	<p>Gaussian Randomized Projection - RMSE - Vehicle_Silhouette</p>  <p>Here the major drop in RMSE (from 27500 to 4900 approx.) happened till dimensions=12. The RMSE reduction slowed after that giving a flatter trend. So, we shall choose dimensions=12 as the optimal.</p> <p><u>Re-running K-Means:</u> Using the Elbow Method, K-means shows the number of optimal clusters k=2. The elbow method conforms with the highest Silhouette Score of 0.6952. So, we shall select k=2 as optimal cluster size.</p> <p><u>Re-Running EM (GMM):</u> Based on the lowest BIC score of 59362.74 the optimal number of clusters has now increased to k=6.</p>

Performance of Clustering Before/After Randomized Projection:

Optimal cluster selections are highlighted in Blue.

	Dataset A: Breast Cancer		Dataset B: Vehicle Silhouette	
	K-Means Silhouette Score	EM-GMM BIC Score	K-Means Silhouette Score	EM-GMM BIC Score
Before GRP	0.3434 (K=2) AMI Score = 0.5318	6174.7942 (K=2) AMI Score = 0.6538	0.3896 (K=2) AMI Score = 0.1225	3668.4657 (K=7) AMI Score = 0.2772
After GRP	0.6992 (K=2) AMI Score = 0.4600	16140.6673 (K=5) AMI Score = 0.3706 16638.8902 (K=2) AMI Score = 0.6431	0.6952 (K=2) AMI Score = 0.1363	59362.7424 (K=6) AMI Score = 0.3458 59377.2081 (K=7) AMI Score = 0.3403

Dataset A:

For the k-Means clustering the Silhouette coeff. almost doubled after performing GRP. This shows that the cluster quality for k-Means has greatly improved after applying GRP. Though there has been reduction in alignment with the ground-truth labels.

EM-GMM the lowest BIC score after dimensionality reduction has actually increased more than double This show that after applying GRP the quality of clusters has degraded. Also, the AMI scores are halved. So, overall the clustering has declined.

Dataset B:

For the k-Means clustering the Silhouette coeff. almost doubled after performing GRP, also AMI has increased which is a very good indicator for the quality of the clusters.

For EM, the quality of clusters has degraded as the BIC scores nearly doubled though the AMI looks to have increased showing a better alignment with the ground-truth labels.

Feature Selection (SelectKBest):

I shall be using SelectKBest algorithm as our feature selection technique here and use “chi2” as the scoring function due to non-negative feature values for our classification dataset.

We shall select the top 50% of the available features from both the datasets based on scoring and then apply clustering. There-after we shall analyse the cluster performance.

Dataset A: Breast Cancer	Dataset B: Vehicle Silhouette																																																																														
Selecting 50% (i.e., 15) of best scoring features.	Selecting 50% (i.e., 9) of best scoring features.																																																																														
<table><tr><th></th><th>Features</th><th>Score</th></tr><tr><td>23</td><td>area_worst</td><td>112598.431564</td></tr><tr><td>3</td><td>area_mean</td><td>53991.655924</td></tr><tr><td>13</td><td>area_se</td><td>8758.504705</td></tr><tr><td>22</td><td>perimeter_worst</td><td>3665.035416</td></tr><tr><td>2</td><td>perimeter_mean</td><td>2011.102864</td></tr><tr><td>20</td><td>radius_worst</td><td>491.689157</td></tr><tr><td>0</td><td>radius_mean</td><td>266.104917</td></tr><tr><td>12</td><td>perimeter_se</td><td>250.571896</td></tr><tr><td>21</td><td>texture_worst</td><td>174.449400</td></tr><tr><td>1</td><td>texture_mean</td><td>93.897508</td></tr><tr><td>26</td><td>concavity_worst</td><td>39.516915</td></tr><tr><td>10</td><td>radius_se</td><td>34.675247</td></tr><tr><td>6</td><td>concavity_mean</td><td>19.712354</td></tr><tr><td>25</td><td>compactness_worst</td><td>19.314922</td></tr><tr><td>27</td><td>concave points_worst</td><td>13.485419</td></tr></table>		Features	Score	23	area_worst	112598.431564	3	area_mean	53991.655924	13	area_se	8758.504705	22	perimeter_worst	3665.035416	2	perimeter_mean	2011.102864	20	radius_worst	491.689157	0	radius_mean	266.104917	12	perimeter_se	250.571896	21	texture_worst	174.449400	1	texture_mean	93.897508	26	concavity_worst	39.516915	10	radius_se	34.675247	6	concavity_mean	19.712354	25	compactness_worst	19.314922	27	concave points_worst	13.485419	<table><tr><th></th><th>Features</th><th>Score</th></tr><tr><td>11</td><td>SCALED_VARIANCE_ALONG_MINOR_AXIS</td><td>12826.234361</td></tr><tr><td>6</td><td>SCATTER_RATIO</td><td>1254.023933</td></tr><tr><td>3</td><td>RADIUS_RATIO</td><td>914.638119</td></tr><tr><td>10</td><td>SCALED_VARIANCE_ALONG_MAJOR_AXIS</td><td>856.125087</td></tr><tr><td>2</td><td>DISTANCE_CIRCULARITY</td><td>503.617020</td></tr><tr><td>15</td><td>KURTOSIS_ABOUT_MINOR_AXIS</td><td>455.300941</td></tr><tr><td>12</td><td>SCALED_RADIUS_OF_GYRATION</td><td>453.874175</td></tr><tr><td>7</td><td>ELONGATEDNESS</td><td>326.845064</td></tr><tr><td>14</td><td>SKEWNESS_ABOUT_MINOR_AXIS</td><td>138.018591</td></tr></table>		Features	Score	11	SCALED_VARIANCE_ALONG_MINOR_AXIS	12826.234361	6	SCATTER_RATIO	1254.023933	3	RADIUS_RATIO	914.638119	10	SCALED_VARIANCE_ALONG_MAJOR_AXIS	856.125087	2	DISTANCE_CIRCULARITY	503.617020	15	KURTOSIS_ABOUT_MINOR_AXIS	455.300941	12	SCALED_RADIUS_OF_GYRATION	453.874175	7	ELONGATEDNESS	326.845064	14	SKEWNESS_ABOUT_MINOR_AXIS	138.018591
	Features	Score																																																																													
23	area_worst	112598.431564																																																																													
3	area_mean	53991.655924																																																																													
13	area_se	8758.504705																																																																													
22	perimeter_worst	3665.035416																																																																													
2	perimeter_mean	2011.102864																																																																													
20	radius_worst	491.689157																																																																													
0	radius_mean	266.104917																																																																													
12	perimeter_se	250.571896																																																																													
21	texture_worst	174.449400																																																																													
1	texture_mean	93.897508																																																																													
26	concavity_worst	39.516915																																																																													
10	radius_se	34.675247																																																																													
6	concavity_mean	19.712354																																																																													
25	compactness_worst	19.314922																																																																													
27	concave points_worst	13.485419																																																																													
	Features	Score																																																																													
11	SCALED_VARIANCE_ALONG_MINOR_AXIS	12826.234361																																																																													
6	SCATTER_RATIO	1254.023933																																																																													
3	RADIUS_RATIO	914.638119																																																																													
10	SCALED_VARIANCE_ALONG_MAJOR_AXIS	856.125087																																																																													
2	DISTANCE_CIRCULARITY	503.617020																																																																													
15	KURTOSIS_ABOUT_MINOR_AXIS	455.300941																																																																													
12	SCALED_RADIUS_OF_GYRATION	453.874175																																																																													
7	ELONGATEDNESS	326.845064																																																																													
14	SKEWNESS_ABOUT_MINOR_AXIS	138.018591																																																																													
<p><u>Re-running K-Means:</u> Using the Elbow Method, K-means shows the number of optimal clusters to be 2 (i.e., k=2). This also conforms with the highest Silhouette_Score of 0.6973. So, <u>we shall select k=2</u> as optimal cluster size.</p> <p><u>Re-Running EM-GMM:</u> Based on the lowest BIC score (18382.116) the optimal number of clusters is now k=8.</p>	<p><u>Re-running K-Means:</u> Using the Elbow Method, K-means shows the number of optimal clusters k=2. The elbow method conforms with the highest Silhouette Score of 0.6822. So, <u>we shall select k=2</u> as optimal cluster size.</p> <p><u>Re-Running EM (GMM):</u> Based on the lowest BIC score of 50261.2285 the optimal number of clusters has now increased to k=8.</p>																																																																														

Performance of Clustering Before/After SelectKBest:

Optimal cluster selections are highlighted in Blue.

	Dataset A: Breast Cancer		Dataset B: Vehicle Silhouette	
	K-Means	EM-GMM	K-Means	EM-GMM

	Silhouette Score	BIC Score	Silhouette Score	BIC Score
Before SelectKBest	0.3434 (K=2) AMI Score = 0.5318	6174.7942 (K=2) AMI Score = 0.6538	0.3896 (K=2) AMI Score = 0.1225	3668.4657 (K=7) AMI Score = 0.2772
After SelectKBest	0.6973 (K=2) AMI Score = 0.4640	18382.1160 (K=8) AMI Score = 0.3503 18454.5942 (K=2) AMI Score = 0.6146	0.6822 (K=2) AMI Score = 0.1363	50261.2285 (K=8) AMI Score = 0.2545 50544.5423 (K=7) AMI Score = 0.2413

Dataset A:

For the k-Means clustering the Silhouette coeff. almost doubled after performing SelectKBest. This shows that the cluster quality for k-Means has greatly improved after applying GRP. Though there has been reduction in alignment with the ground-truth labels.

EM-GMM the lowest BIC score after dimensionality reduction has actually increased more than double This show that after applying GRP the quality of clusters has degraded. Also, the AMI scores are halved. So, overall the clustering has declined.

Dataset B:

For the k-Means clustering the Silhouette coeff. almost doubled after performing GRP, also AMI has increased which is a very good indicator for the quality of the clusters.

For EM, the quality of clusters has degraded as the BIC scores nearly doubled as well as the AMI has dropped. So, overall cluster quality has degraded after applying SelectKBest.

We can observe that there is a similarity between GaussianRandomProjection and SelectKBest in terms of how it affects the overall performance/quality of the clusters.

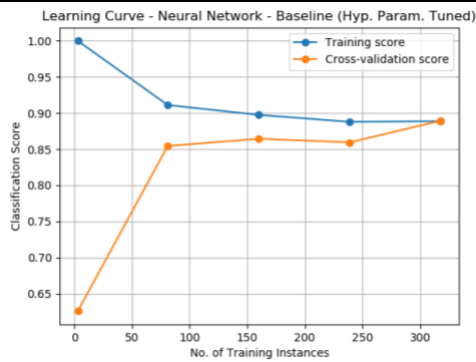
◇ Artificial Neural Network on Baseline and Transformed Dataset:

We shall be doing the following experiments on Neural Network using the Breast Cancer Dataset (Dataset A).

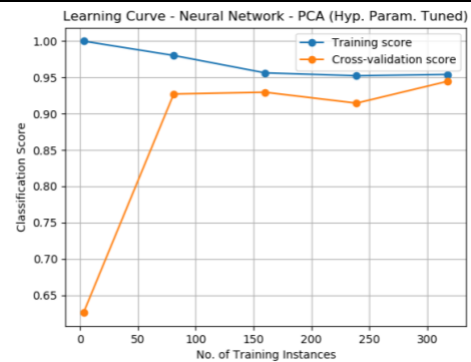
1. Baseline – Neural network ran on “Dataset A” with optimal hyper-parameters as done in assignment 1.
2. Transform “Dataset A” using the following 6 techniques and use the transformed dataset as input to the Neural Network.
 - PCA, ICA, Randomized Projection, Feature Selection (i.e., SelectKBest), K-Means and Expectation Maximization.
3. Perform 5-fold cross-validation with train and test split of the transformed datasets and perform Hyper-parameter tuning same as in Assignment 1 using “Learning Curve” and “Model Complexity Curve”.
4. Capture the Test Accuracy Score, Train and Query time of the optimal models and compare.
5. Due to limitation of space we won’t be detailing on the hyper-parameter tuning process but go straight into the tuned optimal model and discuss on the best performing algorithm
6. We shall be using Neural Network with the same structure as in Assignment 1, i.e., with 2 hidden layers of 4 and 3 nodes respectively and 3000 iterations. The two most crucial hyper-parameters that control the performance of an ANN classifier are it’s Learning Rate and Alpha which shall be tuned to their optimal performance.
7. Breast Cancer dataset is chosen as it would keep the model comparison simpler with Binary Classification and we achieved a very high accuracy score of 97.66% in assignment 1 (baseline) and easier to compare how other transformations exceed or lag in performance.

Below I have showed the performance of the Neural Network along with optimally tuned hyper-parameters and Accuracy Score achieved for each of the transformed datasets.

Baseline (Neural Network on Breast Cancer Dataset)	PCA
learning_rate_init = 0.001; alpha = 10	learning_rate_init = 0.01 ; alpha = 0.1



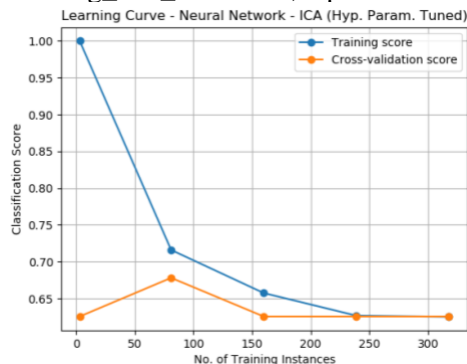
Test Accuracy Score: 97.66 %



Test Accuracy Score: 97.08 %

ICA

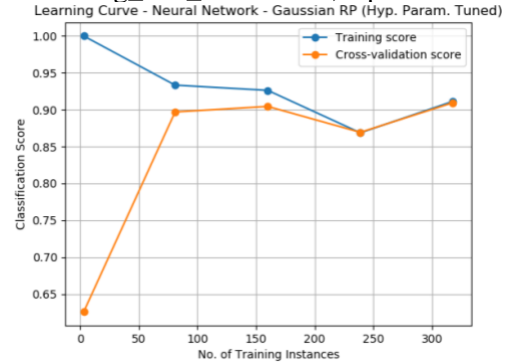
learning_rate_init = 0.1 ; alpha = 0.0001



Test Accuracy Score: 96.49 %

Gaussian Randomized Projection

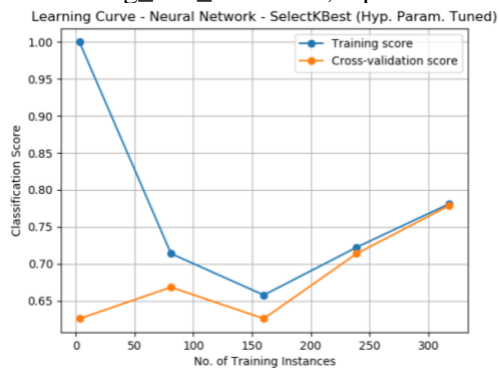
learning_rate_init = 0.01 ; alpha = 100



Test Accuracy Score: 96.49%

SelectKBest

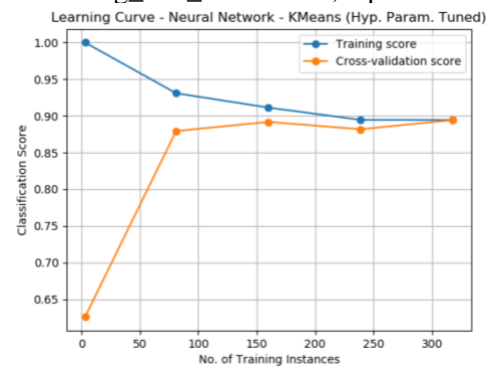
learning_rate_init = 0.1 ; alpha=100



Test Accuracy Score: 95.32 %

K-Means

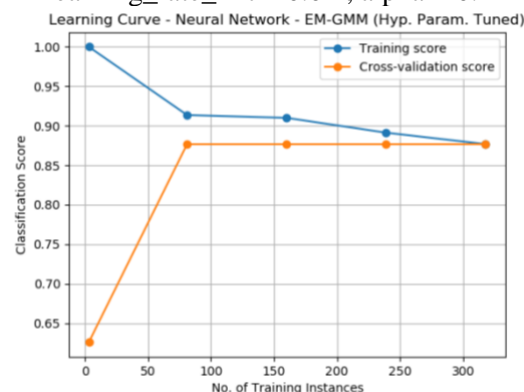
learning_rate_init = 0.01 ; alpha = 10



Test Accuracy Score: 94.74 %

Expectation Maximization (GMM)

learning_rate_init = 0.01 ; alpha = 0.1

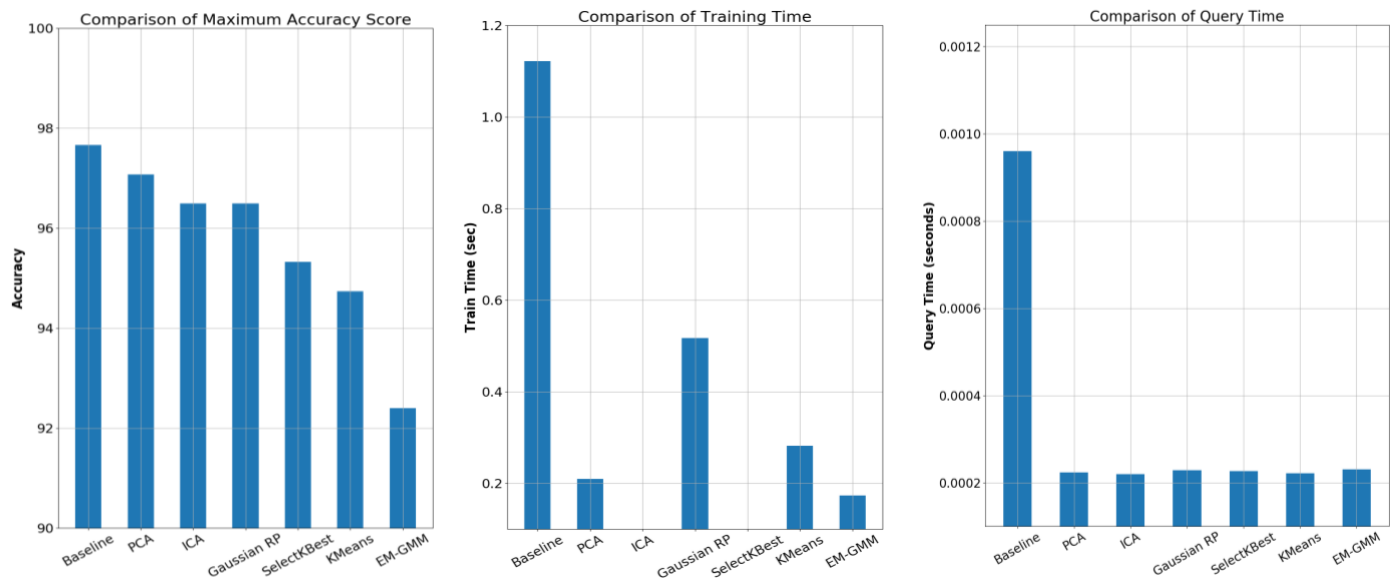


Test Accuracy Score: 92.40 %

Comparison Neural Network performance with various transformed datasets as input:

The accuracy score (on test data) of the Baseline is the highest along with train and query time. The reason could be “information loosing” nature of these Dimensionality Reduction and Clustering algorithm. There can be features with distinguishing characteristics that could be lost by these transformation algorithms. This not only reduces the accuracy of the model but also increases the train time as the Neural Network model takes more time to converge.

Among all the data transformation/selection algorithms, PCA performs is the best. This may show the effectiveness of PCA in understanding the more distinguishing features and preserve them. That’s why we can see that the training time of PCA is also moderate and Query time is almost the same magnitude with little variation than other algorithms.



So, if we do a trade-off between Accuracy, Train Time and Query Time and look for a mode (including baseline model) that maintains a good balance between these three. Then I would say PCA is the best idea in this situation.

Citations (Theory, data and Code referenced from the below sources):

- [1] <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>
- [2] <https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6>
- [3] <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>
- [4] Machine Learning with scikit-learn Quick Start Guide by Kevin Jolly
- [5] [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))
- [6] [https://archive.ics.uci.edu/ml/datasets/Statlog+\(Vehicle+Silhouettes\)](https://archive.ics.uci.edu/ml/datasets/Statlog+(Vehicle+Silhouettes))